植 物 分 类 与 资 源 学 报 2011, 33 (1): 39~45
Plant Diversity and Resources DOI: 10.3724/SP.J.1143.2011.10239

Botanical Literature Goes Global: The Biodiversity Heritage Library

Judith A. WARNEMENT*
(*Botany Libraries*, *Harvard University Herbaria*, 22 Divinity Ave., Cambridge, MA, USA)

Abstract: Scholars in the natural sciences rely on historic literature more than any other branch of science. Yet much of this material has limited global distribution and much of it is available in only a few select libraries. This wealth of knowledge is available only to those few who can gain direct access to significant library collections, a situation that is considered one of the chief impediments to the efficiency of research in the field. Community support and new technologies led to the formation of the Biodiversity Heritage Library. The BHL is an international collaboration of natural history libraries working together to make biodiversity literature available for use by the widest possible audience through open access and sustainable management.

Key words: Biodiversity Heritage Library; Encyclopedia of Life; Botanical libraries; Digital libraries; Taxonomic intelligence; Taxonomic literature

CLC number: G 25 **Document Code**: A **Article ID**: 2095-0845(2011)01-039-07

Background

The biodiversity community is at the forefront of developing international standards and applying new technologies to merge and expand historic datasets with current research, as exemplified by the Biodiversity Heritage Library (BHL) program. The idea for this project started in March, 2005, as scientists, informatics experts, and librarians convened at a session entitled "Libraries and Laboratories" hosted by the Natural History Museum in London to share ideas, goals, and concerns. One outcome was a shared vision to build an integrated digital biodiversity library modeled after Botanicus, Missouri Botanical Garden's digital library. A follow-up organizational meeting was hosted by the Smithsonian Institution Library (SIL) in June of 2006, where librarians from major natural history, botanical garden, and research institutions in the United States and Great Britain were invited to participate in a consortium that would build a global digital collection. All of the participants agreed to move forward, with the Missouri Botanical Garden agreeing to support the development of the technical infrastructure. By February of 2007 a formal organizational meeting was hosted by Harvard's Museum of Comparative Zoology, where the governance structure, operational plans, and working committees for the BHL were formed. Charter members included natural history museum libraries (American Museum of Natural History; Field Museum; Natural History Museum, London; and Smithsonian Institution), botanical garden libraries (Missouri Botanical Garden; New York Botanical Garden; and Royal Botanic Gardens, Kew), as well as academic and research libraries (Harvard University's Botany Libraries; Ernst Mayr Library of the Museum of Comparative Zoology; and Marine Biological Laboratory/Woods Hole Oceanographic Institution Library). Libraries representing the Academy of Natural Sciences (Philadelphia) and California Academy of Sciences joined in 2008. The Biodiversity Heritage Library portal (www.biodiversitylibrary.org; see Fig. 1) was officially

* Author for correspondence; E-mail: warnemen@oeb.harvard.edu; phone (617) 495-2366; Fax (617) 495-8654
Received date: 2010-12-13, Accepted date: 2010-12-30

launched in May, 2007, with Botanicus as the underpinning of a rapidly expanding biodiversity library.

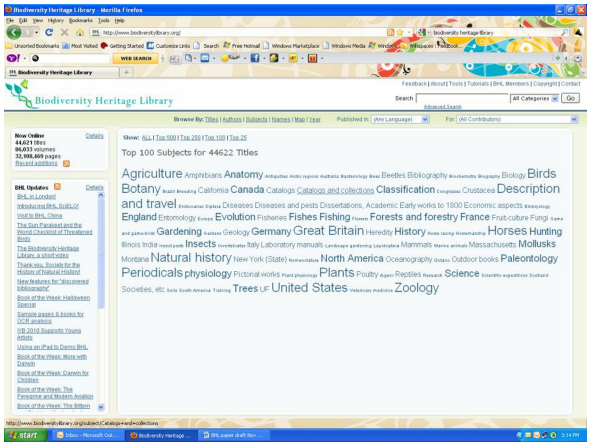

Fig. 1 The Biodiveristy Heritage Library portal

Materials and methods

Partners in the Biodiversity Heritage Library (BHL) program are working together to digitize the published literature of biodiversity held in their respective collections, and to make that literature available for open access as a part of a global "biodiversity commons". The BHL program is the scanning and digitization component of the Encyclopedia of Life (EOL) (www. eol. org), Harvard University's Professor Edward O. Wilson's vision to create a web page for every species of the earth's biota. Another key collaborator is the Internet Archive (IA) (www. archive. org), which is dedicated to "universal access to human knowledge" and provides most BHL partners with low cost mass scanning, archival storage of files, image processing, and technology development. Scanning facilities have been opened in London, Boston, New York, and Washington, D. C. to assist with the BHL and other scanning projects. The IA also allows the BHL to "ingest" other natural history content contributed by non-BHL partners like the California Digital Library, the University of Illinois at Urbana-Champaign, the University of Toronto, and the Boston Library Consortium. The partnership enriches the BHL collection and leverages limited scanning dollars.

The Biodiversity Heritage Library is not a legal entity, but a federation of libraries bound by memoranda of understanding. Members have signed agreements and the library directors represent their respective institutions on an Institutional Council. An elected Executive Committee conducts routine business, and works closely with the three salaried positions that include an executive director, a technical director, and a collections coordinator. There are weekly teleconferences scheduled by the Executive Committee, monthly calls scheduled with the Institutional Council, and there is one face-to-face meeting held each year. These two groups oversee policy and funding decisions, while the details are managed by a variety of broadly representative committees. The scanning staff members teleconference weekly by phone and have been instrumental in developing the tools that manage bidding, workflow, and quality control protocols. A collections committee monitors the overall cohesiveness of BHL content, refines ingest criteria, and reviews all collections-related issues. An active technical committee designs all aspects of the BHL global infrastructure, explores and engages in partnerships that will advance the project's mission. The project is supported by the Encyclopedia of Life budget with grants from the John D. and Catherine T. MacArthur Foundation and the Alfred P. Sloan Foundation, funds from EOL's five anchor institutions, the partner institutions, and other grants.

The BHL consortium is working with the global taxonomic community, rights holders and other interested parties to ensure that this biodiversity heritage is available to all and contributes to the International Convention on Biological Diversity (CBD) and the Global Biodiversity Information Facility (GBIF). "Taxonomic intelligence" is the inclusion of taxonomic practices, skills and knowledge within informatics services to manage information about organisms. Dubbed the Universal Biological Indexer and Organizer, or uBio, BHL is using a sophisticated algorithm to locate likely name strings in OCR text,

has "discovered" 10.7 million name strings in NameBank (Fig. 2) (www.ubio.org/index.php?pagename=namebank), and serves as a name thesaurus. The link between the name service and the BHL collection creates a powerful new tool for scholars.

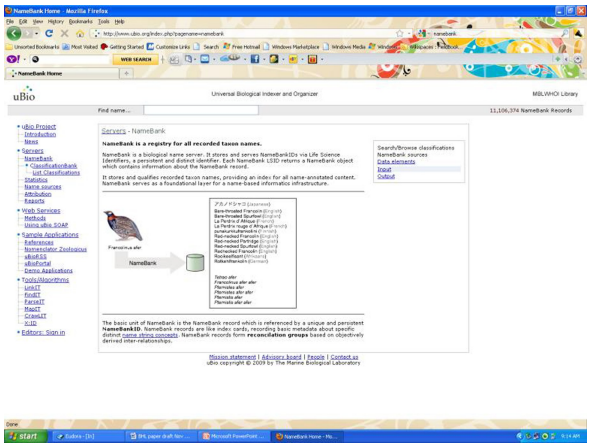

Fig. 2　NameBank portal

Systematists and taxonomists need access to the historic literature to support current research. The cited half-life of publications in taxonomy and the "decay rate" are longer than in any other scientific discipline so the "current" biodiversity literature spans more than 250 years. At the outset of the project, the BHL needed to calculate the scope of the biodiversity domain. OCLC (www.oclc.org/us/en/default.htm), the major international library utility, supported the BHL by merging all of the partners' catalog records into a database so OCLC's collection analysis tool could be used to profile the overall collection. The outcome showed that biodiversity literature is represented by 1.3 million catalog records. More than 800 000 records describe monographs and 40 000 records describe journal titles, with 12 500 records representing current titles. About forty percent of the material was published prior to 1923, generally placing it the public domain. Sixty-three percent of the records were for works in English, and German was the second most frequent language (9%). The BHL's scanning efforts have focused on the pre-1923 content that is not readily available, and yet essential to taxonomists' research.

The technical team developed tools to coordinate scanning efforts and avoid duplication. A merged database of members' serials holdings was created as a "bid list" so that each library can indicate the titles it intends to scan. If problems are discovered, such as missing volumes, or pages, or illustrations, then a call goes out to other libraries to scan those volumes. Monographs are selected by each library along subject areas, and the BHL collection is checked prior to scanning to avoid duplication. All items are barcoded and shipping manifests are created using a tool called WonderFetch (biodiversitylibrary.blogspot.com/2008/06/wonderfetchtm-ia-metaxml-fields.htm). The partner libraries can populate fields with data that would not normally be populated as part of the standard IA process, and then store those values alongside each scanned item in the IA repository. The impetus for implementing WonderFetch was not just to automate the inclusion of essential data elements like the volume and issue information for serials, but to also capture due diligence, rights, and licensing information related to each item. Partner libraries underwrite all of the costs associated with identifying, processing, and shipping materials, and BHL grants support the costs associated with scanning and digital processing.

Results and Discussion

The BHL portal currently offers more than 44 000 titles represented by nearly 86 000 volumes delivering more than 32 million pages of content. Users can search by simply browsing by author, title, or subject, or can use the novel language, year of publication, and source map options (Fig. 3). More refined searches can be achieved by using the search box that allows the user to search for a specific author, title, subject, or species names. It is the delivery of search results that is unique to BHL. Species names results are delivered as a bibliography that cites the source title, author, date, and pages, and includes a link to the NameBank record. The

pages of any volume selected are automatically scanned by the uBio search feature for species names. The results appear in the "Names on this page" box on the lower left-hand corner of the screen (Fig. 4). Links to EOL species pages are highlighted, and clicking on any of the discovered names will generate a species name bibliography. Searchers can click on any name in the uBio box to see a bibliography of all other occurrences of the name in the entire portal. For example, selecting *Rhododendron indicum* will generate the bibliography shown in Fig. 5 that includes links to all source materials.

The "Download/About this book" tab appears (Fig. 6) when a title is displayed. Users are able to download the bibliographic record, selected pages, images, or the entire volume. The menu also features PDF or OCR download options, and links to views via other portals. In order to download selected pages, users supply an email address, and a citation for the request, and then select up to one hundred pages. These documents are retained when appropriate metadata is provided and are made available to other users through CiteBank (citebank. org). Citebank (Fig. 7) is still under development, but in addition to saving the BHL-selected documents, it is intended to provide robust search and browse capabilities to biodiversity publications stored in multiple international repositories and aggregate content from as many systems as possible, so that biodiversity

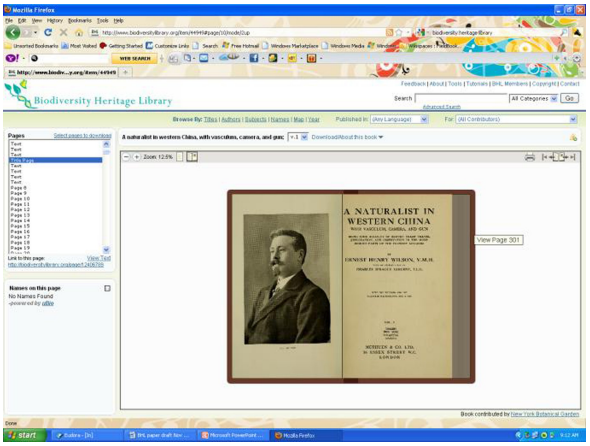

Fig. 3 Portrait and title page in Ernest H. Wilsons *A Naturalist in Western China*. (London: Methuen, 1913)

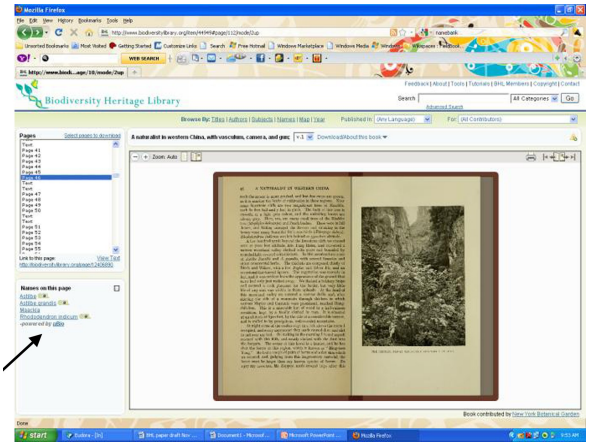

Fig. 4 Species names discovered by uBio appear in box in lower left corner. Note the links to EOL species pages

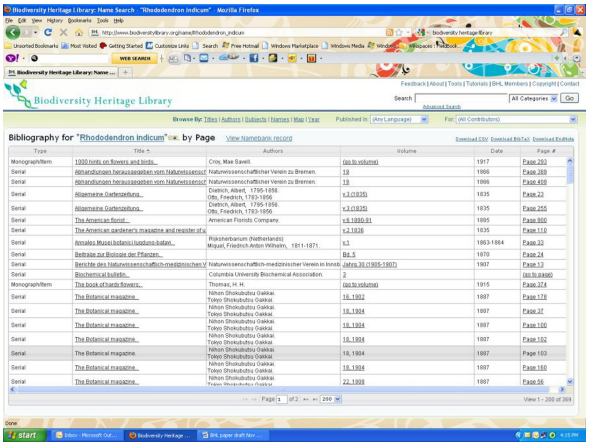

Fig. 5 uBio generated bibliography for *Rhododendron indicum*

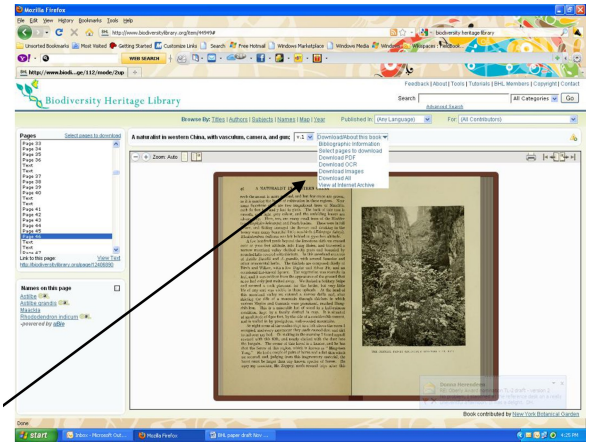

Fig. 6 BHL Download/About this book tab

researchers have a single point of access to published materials. CiteBank is also intended to provide a storage platform for articles and documents that are digitized, but not yet online and offer a common system for researchers to share specialized bibliographies. Users will be able to upload, edit, and share their own personal lists of references and citations, and these references will be linked to scanned content in the Biodiversity Heritage Library portal. In addition, BHL offers to scan professional societies' publications or other publishers content at BHL's expense and integrate the content into the BHL. Agreements with nearly forty publishers have added about one hundred titles to the collection.

The Biodiversity Heritage Library wiki (Fig. 8) (biodivlib. wikispaces. com) presents a wealth of information about the project, detailed instructions and tutorials for using the various features, and lists of members and BHL staff. Developers' tools are described and documented, and the BHL's opt-in copyright feature is explained. The BHL also uses popular social media tools to connect with the public, including Twitter, Facebook, and the BHL blog (biodiversitylibrary. blogspot. com). Each site attracts and supports a varied community of scientists and the general public.

Interest and support in the BHL has grown at an astonishing rate. In less than five years the BHL has grown into an international partnership that mirrors the global nature of biodiversity research. Formal BHL agreements are in place in Europe, China, and Australia, and there is strong interest in South America and Egypt. In **Europe**, colleagues at twenty-eight European institutions have obtained funding from the European Union eContentplus Program to establish a BHL-Europe (www. bhl-europe. eu), which is developing the technical infrastructure and tools to deliver content from many scanning projects throughout the continent. In the United Kingdom work is also proceeding via the BHL and Europeana (Fig. 9) (www. europeana. eu/portal). In **China**, the Chinese Academy of Sciences supports BHL-China (Fig. 10) (www. bhl-china. org/cms/en), and the Internet Archive installed a scanner in Beijing in the summer of 2010 to help build the BHL-China collection. In **Australia**, The Atlas of Living Australia, funded by the Australian government's National Collaborative Research Infrastructure Strategy program joined BHL in June 2010 (www. ala. org. au). Additional partnerships, policies, tools, and tutorials are being explored and developed to refine the BHL to increasingly extend its global reach.

A great deal has been achieved through conventional mass scanning technologies and practices, but a significant portion of early biodiversity literature is quite rare and valuable, sometimes fragile, and often the book is too large or has folded maps or illustrations that do not fit on conventional scanning beds. A planning grant, *Retooling Special Collections in the Age of Mass Digitization*, awarded by the Institute of Museum and Library Services (IMLS) in 2008 allowed BHL partners to identify and develop a cost-effective and efficient large-scale digitization workflow and to explore ways to enhance metadata for library materials that are designated as "special collections." The group held a series of meetings, communicated by email, and established a wiki to record meetings, track progress, and share documents about costs, statistics and workflows, and small-scale scanning tests. The report included extensive cost analyses and recommendations for equipment configurations to scan rare and oversized materials.

BHL partners are also exploring ways to introduce other essential content to the BHL portal. Collectors' field notes, plant lists, and diaries often hold important information that supplements content found on specimen labels and published accounts. Access to this primary source material is even more problematic to scholars because most archival collections, if catalogued, are not described in very fine detail. The United States National Herbarium and Smithsonian Institution Archives have received a Cataloging Hidden Special Collections and Archives

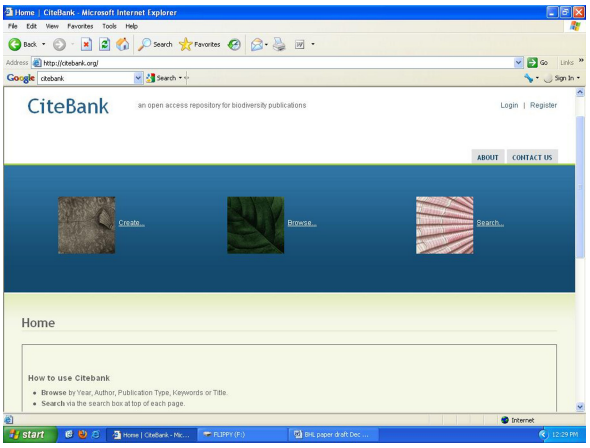

Fig. 7 CiteBank homepage

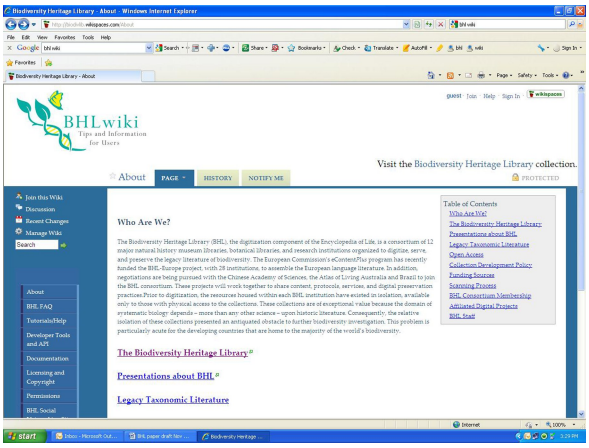

Fig. 8 BHL wiki

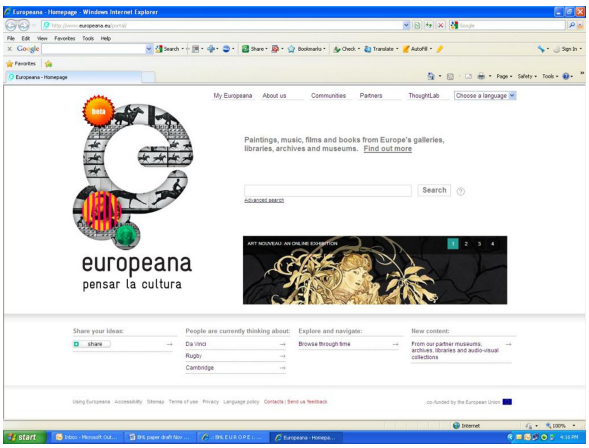

Fig. 9 Europeana portal

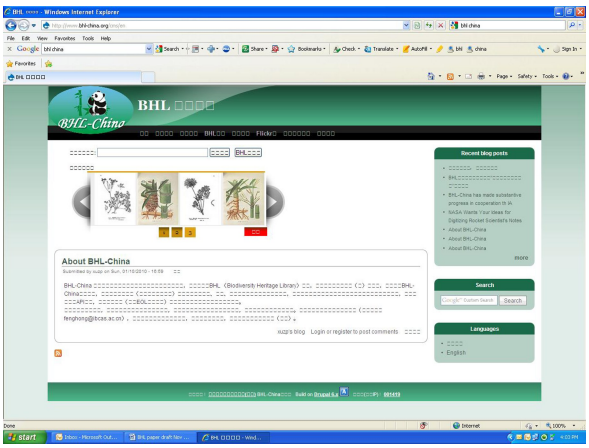

Fig. 10 BHL-China portal

Grant, from the Council on Library and Information Resources (CLIR) to catalog all the field books, unpublished journals, loose notes, and sketches that document field research related to all disciplines of biology. The grant, *Exposing Biodiversity Field Books and Original Expedition Journals at the Smithsonian Institution*, will also will build a cataloging tool to and create a central repository so that other institutions can contribute their holdings. The enhanced level of description will improve access to these important research materials that are frequently difficult to discover and access remotely.

Several BHL partners have been awarded an IMLS grant as a companion grant to the Smithsonian's CLIR proposal. *Connecting Content: A Collaboration to Link Field Notes to Specimens and Published Literature* will develop a system for integrating biological researchers' field and specimen notes with museum specimens and related electronically published literature. The enhanced and integrated access to biological data will serve a wide variety of users, and will connect to other ongoing projects such as the Biodiversity Heritage Library.

The Biodiversity Heritage Library will soon benefit from another new collaboration. The International Association of Plant Taxonomists (IAPT) has given their permission to rescan and integrate with the BHL the monumental fifteen volume botanical bibliography, *Taxonomic Literature*, 2nd ed. (TL-2). The Smithsonian Institution Library has been awarded an Atherton Seidell Grant to accomplish the scanning and design the schema. The BHL envisions

a dynamically linked "TL-3" that will connect citations to published references and allow for corrections and the addition of new and expanded content.

The Biodiversity Heritage Library has achieved remarkable success in its relatively short existence. The partners have demonstrated that independent and geographically dispersed institutions can collaborate effectively, and have proven their ability to generate significant financial support. The technical accomplishments by a small team of talented and dedicated informatics specialists and the efficient and collegial intra-institutional working groups are apparent in the array of tools and services currently delivered and under development via the various BHL interfaces. On June 27, 2010, the American Library Association's Association for Library Collections & Technical Services (ALCTS) awarded their Outstanding Collaboration Citation to the BHL in recognition of their outstanding collaborative partnership.

The project has generated excitement in the international community and many opportunities to develop new partnerships and sources of funding. Society journal publishers are enthusiastic about participation in the BHL opt-in copyright model. The portal has recorded nearly 1.5 million visits since January of 2008, the taxonomic intelligence tool is highly effective, and there are high levels of OCR accuracy in late 19th and 20th century printing. However, the Biodiversity Heritage Library faces many challenges in the near future. Initial sources of funding end in 2012, and a plan for financial and digital sustainability must be formulated. The rapid international expansion of BHL presents new governance issues, increases the need for clear and focused standards, and strategies to avoid duplication of effort. BHL is working to ensure the technical infrastructure for delivering and preserving content through digitization and retrospective ingestion, as well as the ability to continue to deliver new services as needed by the community.

Acknowledgements: The author wishes to acknowledge all of the BHL partners for their collegiality and dedication, and Dr. Jinshuang Ma and the Shanghai Chenshan Botanical Garden and Plant Science Research Center for their support.

References:

Gwinn NE, Rinaldo C, 2009. The Biodiversity Heritage Library: Sharing Biodiversity literature with the world [J]. *IFLA Journal*, **35**: 25—34 DOI: 10.1177/0340035208102032

International Association of Aquatic and Marine Science Libraries and Information Centers Conference (35th: 2009: Brugge, Belgium). IAMSLIC Proceedings http://hdl.handle.net/1912/3787

Rinaldo C, 2009. The Biodiversity Heritage Library: exposing the taxonomic literature [J]. *Journal of Agricultural & Food Information*, **10**: 259—265 DOI: 10.1080/10496500903014669

Rinaldo C, Norton C, 2010. The Biodiversity Heritage Library: an expanding international collaboration